*Electronic Journal of Applied Statistical Analysis*
EJASA (2013), Electron. J. App. Stat. Anal., Vol. 6, Issue 1, 110 – 117
e-ISSN 2070-5948, DOI 10.1285/i20705948v6n1p110


# ON THE EXPECTED DIFFERENCE BETWEEN MEAN AND MEDIAN

**Robin de Nijs*, Thomas Levin Klausen**

*Department of Clinical Physiology, Nuclear Medicine and PET, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark*

Received 19 August 2011; Accepted 31 May 2012
Available online 26 April 2013

***Abstract***: *Mean and median are both estimators of the central value of statistical distributions. For a given sampled continuous probability distribution function the equations for the influence of one sample are used to derive the known expressions for the standard deviation of mean and median and the novel expression for the expectation value of the squared differerence between mean and median. The usefulness of this result is illustrated on a dataset, where the difference between mean and median serves as an outlier detection and as a tool for deciding whether to use the mean or median as an estimator.*

***Keywords***: *Mean, median, standard deviation, efficiency, expectation, outlier.*

## 1. Introduction

Both mean (average) and median are commonly used as estimators for the "true" value of a sampled quantity. The uncertainty of these estimators is given by their standard deviation (SD). For normal distributions the SD of the mean is lower than the SD of the median. This means that in this case the mean is a better estimator in terms of uncertainty. On the other hand, if outliers are present, the SD of the mean may become very large, while the SD of the median remains more or less the same. In the case of an outlier the mean itself is potentially changed drastically, while the median remains close to the central value of the original statistical distribution. These are well-known properties of the mean and median, and it is said, that the median is a more robust estimator of the central value. The question, that remains, is what the numerical difference between mean and median is for a given dataset. In this note an equation for the (squared) expected difference between mean and median is derived and it is shown how this can be used as a tool for analyzing a dataset.

* Email: robin.de.nijs@rh.regionh.dk

## 2. Materials and Methods

SD is a measure for the uncertainty or statistical noise. The SD of the mean $\bar{x}$ is just the SD of the population divided by the square root of the number of samples. The symbol *s* will be used to indicate the empirical SD of a set of samples, while the symbol σ will be used for the SD of a continuous probability distribution. The SD of the median can be approximated with [1, 3]:

$$s^2(\tilde{x}) = \alpha_n^2 \cdot s^2(\bar{x}) \approx \frac{1}{4np^2(\tilde{x})} \quad , \tag{1}$$

where *n* is the number of samples, *p* the underlying continuous probability density function, $\tilde{x}$ the median, and α a constant depending on the number of samples and related to the relative efficiency of the median (=$\alpha^{-2}$). The (asymptotic) approximation holds for large *n*, but the exact values of the efficiency can be calculated, see Table 1 and 2. For instance, the underestimation of the variance of the median is less than 10 %, and consequently less than 5 % for α, for *n*>4 for normal distributions and odd values of *n* [8]. The asymptotic efficiency $\alpha_\infty$ is $\sqrt{\pi/2} \approx 1.25$ for normal distributions with width σ, which have $p^{-1}(\tilde{x}) = \sigma\sqrt{2\pi}$ and $n \cdot s^2(\bar{x}) = \sigma^2$. For uniform distributions with width *d* the asymptotic value is $\alpha_\infty = \sqrt{3} \approx 1.73$ with $p(\tilde{x}) = d^{-1}$ and $n \cdot s^2(\bar{x}) = \mathrm{d}^2/12$. This means that for these distributions the uncertainty in the median is bigger than in the mean. For symmetric (non-skewed) distributions the expectation value of mean and median is the same. So there should be no difference between these two apart from uncertainty. However, this changes when so-called outliers are added to the dataset. The median minimizes the total absolute deviation, while the mean minimizes the total squared deviation. This means that the median is less sensitive (more robust) to outliers, but on the other hand has a larger uncertainty in the absence of outliers.

### *2.1 The influence of one sample*

The difference between mean and median is nicely illustrated by the influence of one sample $x_i$ of a data set with *n* samples and mean $\bar{x}$. The mean is shifted (shift is indicated by Δ) by:

$$\Delta\bar{x}_i = \frac{x_i - \bar{x}}{n} \equiv \frac{\Delta x_i}{n}. \tag{2}$$

The influence of one sample depends both on the number of samples and the sample value.
For a large number of samples *n* and an assumed underlying continuous probability distribution function *p(x)* the shift in the median $\tilde{x}$ caused by one sample of the dataset can be calculated by sampling the underlying continuous probability distribution function (pdf) around the median in such a way that the area under the pdf with width *w* corresponds to one sample. This means that the integral of the continuous pdf around the median times the number of samples equals one. The width *w* is equal to the sampling distance yielding a sampling density of one around the median. This corresponds to that the median is shifted half a position in the ordered list of samples. The shift in the median is then given by:

$$\Delta\tilde{x}_i = \frac{w}{2} \approx \frac{1}{2np(\tilde{x})}, \tag{3}$$

and shifted in the opposite direction for samples with a value lower than the median. Alternatively equation (3) can be derived by the analysis of the influence of one sample on the cumulative probability distribution. The median is defined by the value of $x$ where the cumulative continuous distribution function $D$ is 1/2, which states that exactly half of the samples have a lower value than the median, see e.g. p.45 [2]. The influence of one sample on the value of $D$ is $1/2n$. This gives the possibility to approximate the influence of one sample on the median for large $n$ with:

$$\frac{dD(\tilde{x})}{dx} \cdot \Delta\tilde{x} \approx \frac{1}{2n} \qquad \longrightarrow \qquad \Delta\tilde{x} = \frac{1}{2np(\tilde{x})}, \tag{4}$$

where the derivative of the cumulative probability distribution is replaced by the probability density function.
The shift in the median does not depend on the value of the extra sample, which is an appealing property in the case of an outlier. Only the position of the median in the ordered list of samples is shifted half a position. In the case of an even number of outliers equally added above and below the true median value the shift is even cancelled.
The probability function of the median itself is normally distributed for any probability function of the samples. Equation (2) and (3) can be used for calculation of the standard deviations of mean and median. The standard deviation of the mean $s(\bar{x})$ for a set of samples given by:

$$s^2(\bar{x}) = \sum_i (\Delta\bar{x}_i)^2 = \frac{<(\Delta x)^2>}{n} = \frac{s^2}{n}, \tag{5}$$

and the standard deviation of the median $s(\tilde{x})$ for a set of samples is given by:

$$s^2(\tilde{x}) = \alpha_n^2 \cdot s^2(\bar{x}) = \sum_i (\Delta\tilde{x}_i)^2 \approx \frac{1}{4np^2(\tilde{x})}, \tag{6}$$

which is equivalent to equation (1).

### *2.2 The expected difference between mean and median*

Because the median and mean are strongly correlated, the expected squared difference is not just the sum of their variances (var=SD$^2$). The expected squared difference of mean and median is given by:

$$<(\bar{x}-\tilde{x})^2> = <\bar{x}-\tilde{x}>^2 + \mathrm{var}(\bar{x}-\tilde{x}) = <\bar{x}-\tilde{x}>^2 + \mathrm{var}(\bar{x}) + \mathrm{var}(\tilde{x}) - 2\cdot\mathrm{cov}(\bar{x},\tilde{x}). \tag{7}$$

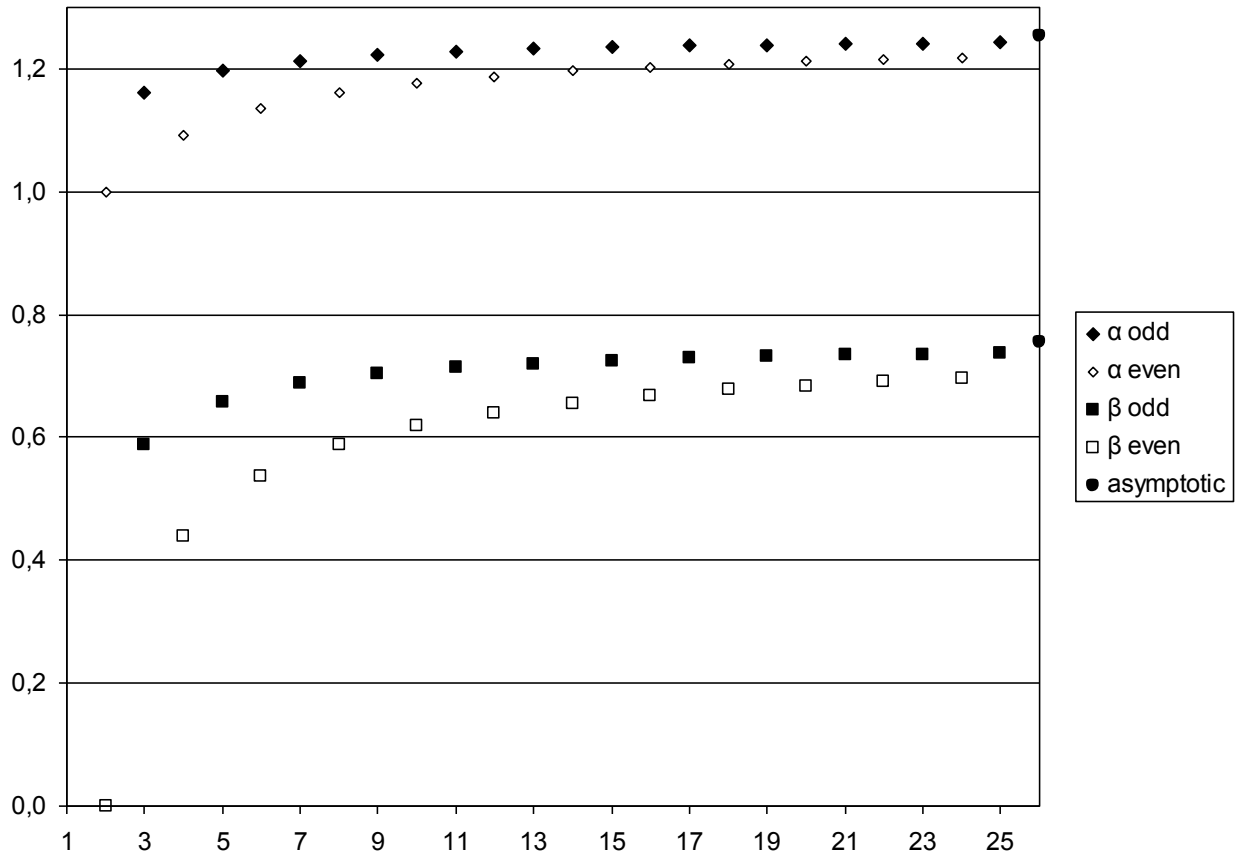


**Figure 1. Values for α and β as a function of the number of samples for a normal distribution.**

The variances of mean and median are given by equations (5) and (6). The covariance $\mathrm{cov}(\bar{x},\tilde{x})$ between mean and median can be calculated for a set of samples with:

$$\mathrm{cov}(\bar{x},\tilde{x}) = \sum_i \Delta\bar{x}_i \Delta\tilde{x}_i \approx \frac{<|x|>}{2np(\tilde{x})}, \tag{8}$$

where $\Delta\bar{x}_i$ and $\Delta\tilde{x}_i$ are defined and approximated with the asymptotic value by equations (2) and (3). For distributions with the same expectation value for mean and median the expectation value of the squared difference between mean and median for a set of samples is given by:

$$<(\bar{x}-\tilde{x})^2> = \beta_n^2 \cdot s^2(\bar{x}) = \sum_i (\Delta\bar{x}_i - \Delta\tilde{x}_i)^2 \approx s^2(\bar{x}) + \frac{1}{4np^2(\tilde{x})} - \frac{<|x|>}{np(\tilde{x})}, \tag{9}$$

where β is a constant related to the relative efficiency. Here the variance is replaced by the squared SD of the dataset. For normal distributions $\beta_\infty$ is $\sqrt{\pi/2-1} \approx 0.76$ and for uniform distributions $\beta_\infty$ equals one. So the expectation value of the squared difference between mean and median is of the order of the variance of the mean for both normal and uniform distributions.

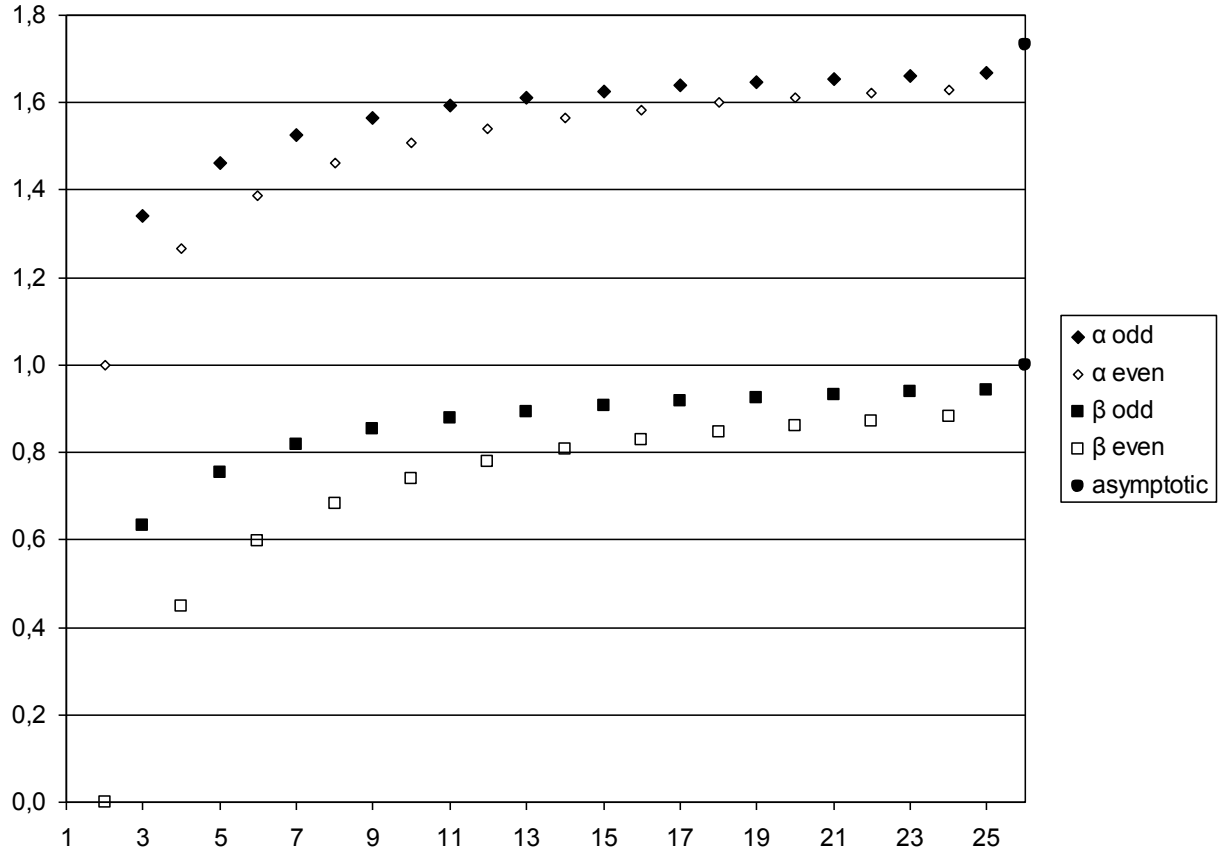


**Figure 2. Values for α and β as a function of the number of samples for a uniform distribution.**

From Chebyshev's inequality it can be derived that the difference between mean and median of a continuous probability density function is less than the standard deviation σ of the underlying distribution [4]. This imposes an upper bound of $n$ on $\beta^2$.

The values α and β are calculated as a function of $n$ for a normal distribution in Table 1 and a uniform distribution in Table 2 with a statistical simulation performed in Matlab 7.9, The Mathworks Inc., Natick, Massachusetts, USA. Figure 1 and 2 show the results from Table 1 and 2 in a graph. In a given dataset the SD of the median and the expectation value of the squared difference between mean and median might be approximated by assuming a certain statistical distribution, calculating the standard deviation of the mean and applying the appropriate value for α and β. The calculated value of the expectation value of the squared difference between mean and median can be compared with the observed difference between mean and median and statistically tested. Assuming a normal distribution of the difference between mean and median the null-hypothesis (no outliers) can be tested with a so-called Z-test, yielding a p-value indicating the statistical significance.

An alternative statistical test for the detection of outliers could be the chi squared test. In this test the observed variance is compared to the known expected variance. Since the variance in general is not known beforehand this test cannot be applied here.

**Table 1. Values for α and β for a normal distribution, calculated by statistical simulation. A graph of the values for α and β can be found in Figure 1.**

| $k$ | $\alpha_{2k}$ | $\alpha_{2k+1}$ | $\beta_{2k}$ | $\beta_{2k+1}$ |
|---|---|---|---|---|
| 1 | 1.0000 | 1.1602 | 0.0000 | 0.5882 |
| 2 | 1.0922 | 1.1976 | 0.4391 | 0.6589 |
| 3 | 1.1351 | 1.2137 | 0.5371 | 0.6878 |
| 4 | 1.1600 | 1.2227 | 0.5878 | 0.7035 |
| 5 | 1.1761 | 1.2283 | 0.6191 | 0.7133 |
| 6 | 1.1875 | 1.2322 | 0.6404 | 0.7200 |
| 7 | 1.1960 | 1.2350 | 0.6560 | 0.7249 |
| 8 | 1.2025 | 1.2372 | 0.6678 | 0.7285 |
| 9 | 1.2077 | 1.2390 | 0.6771 | 0.7314 |
| 10 | 1.2119 | 1.2404 | 0.6846 | 0.7338 |
| 50 | 1.2445 | 1.2506 | 0.7408 | 0.7511 |
| ∞ | 1.2533 | 1.2533 | 0.7555 | 0.7555 |

**Table 2. Values for α and β for a uniform distribution, calculated by statistical simulation. A graph of the values for α and β can be found in Figure 2.**

| $k$ | $\alpha_{2k}$ | $\alpha_{2k+1}$ | $\beta_{2k}$ | $\beta_{2k+1}$ |
|---|---|---|---|---|
| 1 | 1.0000 | 1.3416 | 0.0000 | 0.6325 |
| 2 | 1.2649 | 1.4638 | 0.4472 | 0.7559 |
| 3 | 1.3888 | 1.5276 | 0.5976 | 0.8165 |
| 4 | 1.4606 | 1.5667 | 0.6832 | 0.8528 |
| 5 | 1.5076 | 1.5933 | 0.7386 | 0.8771 |
| 6 | 1.5407 | 1.6125 | 0.7774 | 0.8944 |
| 7 | 1.5653 | 1.6270 | 0.8062 | 0.9075 |
| 8 | 1.5843 | 1.6384 | 0.8284 | 0.9177 |
| 9 | 1.5993 | 1.6475 | 0.8460 | 0.9259 |
| 10 | 1.6116 | 1.6550 | 0.8603 | 0.9325 |
| 50 | 1.7065 | 1.7152 | 0.9705 | 0.9854 |
| ∞ | 1.7321 | 1.7321 | 1.0000 | 1.0000 |

However, in the case of Poisson statistics the chi squared test might be applied, since in this case the variance is equal to the mean itself.

## 3. Results

The calculated values for α are in agreement with the values found in literature [5,8,9]. The relative deviation from the asymptotic value of the variance of the median is as expected approximately twice the relative deviation in α. For even numbers of $n$ the convergence the asymptotic value is slower to than for odd values.

In Figure 3 a histogram is shown of measurements of the specific binding ratio (SBR) for dopamine transporters, measured with Iodine-123 SPECT [6], for healthy subjects and for patients (Parkinson disease) with decreased SBR. Each person has two SBRs; a SBR for the left and right part of the brain. In order to demonstrate the difference between mean and median first the statistics of the healthy subject group are calculated and then the statistics of this group with 1 and 2 patients added. The results are shown in Table 3.

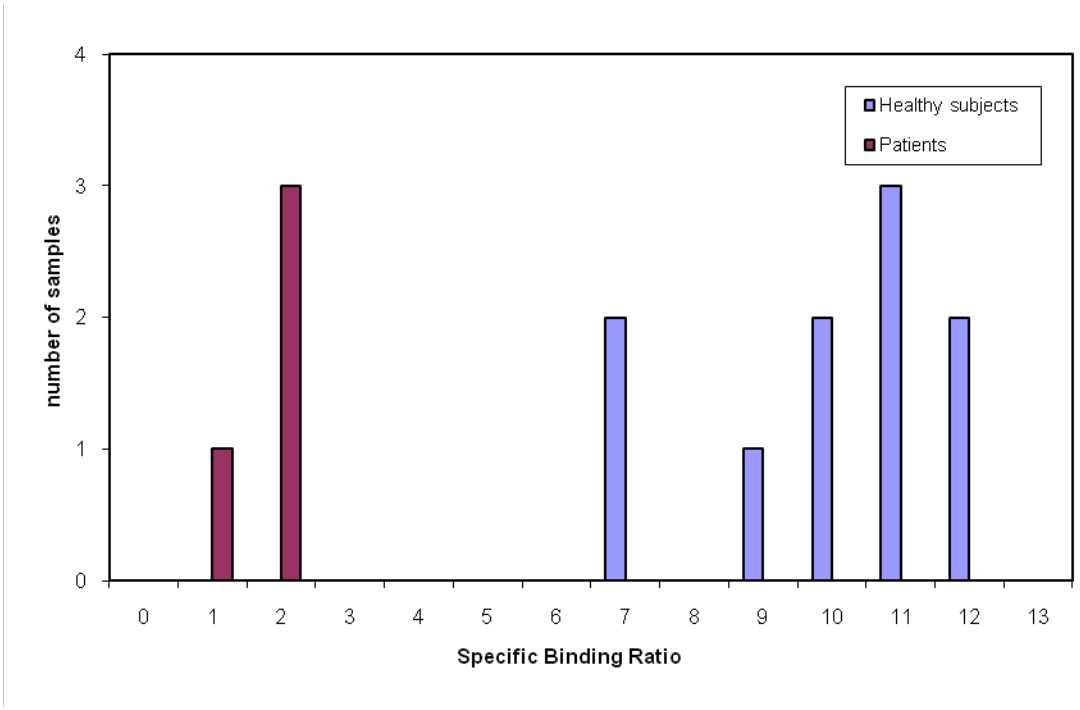


**Figure 3. A histogram of the specific binding ratio (both of the left and right striatum) for dopamine transporter for five healthy subject and two Parkinson patients.**

In the column with 10 samples in Table 3 it can be seen, that the difference between mean and median is of the same order as the SD of the mean. Things change drastically when so-called outliers (the patients in this dataset) are added to the dataset, as illustrated by the last two columns in Table 3.

**Table 3. The mean, median, their standard deviations and their difference for the SBR dataset. Each subject has two samples (left and right striatum). The second column with 10 samples shows the data for the 5 healthy subjects. The third column shows the dataset of the second column with one patient (outlier) added, and the last column shows this dataset with two patients (outliers) added. The last two rows show the z- and p-value for the statistical testing of the observed difference between mean and median against the difference according to equation (9) assuming un underlying normal distribution.**

| **Number of samples** | **10** | **12** | **14** |
|---|---|---|---|
| Mean | 10.4 | 9.0 | 8.0 |
| Median | 10.9 | 10.6 | 10.2 |
| SD of the Mean | 0.56 | 1.06 | 1.10 |
| SD of the Median* | 0.70 | 1.31 | 1.36 |
| Mean-Median, observed | -0.45 | -1.61 | -2.18 |
| Mean-Median* | 0.37 | 0.71 | 0.75 |
| *z* | 1.22 | 2.28 | 2.91 |
| *p*# | 0.222 | 0.022 | 0.004 |

**estimated with equation (6) and (9) assuming an underlying normal distribution*
*#double sided p-value, assuming the difference between mean and median is normally distributed*

The SD of the mean gets bigger, but most profound is, that while the difference between mean and median in the first column was of the same order as the SD of the mean, the difference between mean and median is drastically increased and becomes clearly larger than the

(empirical) SD of the mean. The calculated p-values indicate that this is statistically significant and that the null-hypothesis (no outliers) needs to be rejected in the last two cases.

## 4. Conclusions

Both mean and median as well as their difference are useful tools for characterizing acquired datasets. In practice both mean and median can be calculated, and their difference can reveal unwanted properties of the acquired dataset. Outliers and asymmetrical (skewed) probability distributions will result in a big difference between mean and median. If the difference between mean and median is limited the mean can be used as the least noise sensitive estimator. Mean and median are not only useful for scalar quantities, but also for the analysis of signals [10] and images [7], where mean filtering (averaging) and median filtering can be applied.
In this statistical note the known result of the variance of mean and median for large sample sizes is shown, and the novel result for the expectation value of the squared difference between mean and median is derived and presented. This gives an opportunity to investigate datasets by looking at the difference between mean and median and comparing it to the uncertainty in the mean.

## References

[1]. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.
[2]. Hogg, R.V., Craig, A.T. (1995). *Introduction to Mathematical Statistics, 5th ed.* New York: Macmillan.
[3]. Kenney, J.F., Keeping, E.S. (1962). §13.13 The Median, in *Mathematics of Statistics*, part 1, 3rd ed. Princeton, NJ: Van Nostrand, 211-212.
[4]. Mallows, C.L., Richter, D. (1969). Inequalities of Chebyshev type involving conditional expectations. *The Annals of Mathematical Statistics*, 40, 1922-1932.
[5]. Maritz, J.S., Jarrett, R.G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, 73, 194-196.
[6]. de Nijs, R., Holm, S., Thomsen, G., Ziebell, M., Svarer, C. (2010). Experimental determination of the weighting factor for the energy window subtraction based downscatter correction for I-123 in brain SPECT-studies. *Journal of Medical Physics*, 35, 215-222.
[7]. Pratt, W.K. (1978). 12.6 Median Filter, in *Digital Image Processing*, New York: John Wiley & Sons Inc., 330-333.
[8]. Rider, P.R. (1960). Variance of the median of small samples from several special populations. *Journal of the American Statistical Association*, 55, 148-150.
[9]. Sheather, S.J. (1986). A finite sample estimate of the variance of the sample median. *Statistics & Probability Letters*, 4, 337-342.
[10]. Slotboom, J., Nirkko, A., Brekenfeld, C., van Ormondt, D. (2009). Reliability testing of in vivo magnetic resonance spectroscopy (MRS) signals and signal artifact reduction by order statistic filtering. *Measurement Science and Technology*, 20, 104030 (14pp).